Infection and Immunity, September 1998, p. 4305–4312, Vol. 66, No. 9
0019-9567/98/$04.00+0

- Abstract of this Article
- Reprint (PDF) Version of this Article
- Similar articles found in:
  IAI Online
  PubMed
- PubMed Citation
- This Article has been cited by:
  other online articles
- Search Medline for articles by:
  McClelland, M Wilson, R. K.
- Alert me when:
  new articles cite this article
- Download to Citation Manager

Comparison of Sample Sequences of the *Salmonella typhi* Genome to the Sequence of the Complete *Escherichia coli* K-12 Genome

Michael McClelland,[1,*] and Richard K. Wilson[2]

Sidney Kimmel Cancer Center, San Diego, California 92121,[1] and Genome Sequencing Center, Washington University School of Medicine, St. Louis, Missouri 63108[2]

Received 12 December 1997/Returned for modification 24 March 1998/Accepted 4 June 1998

ABSTRACT

- Top
- Abstract
- Introduction
- Materials & Methods
- Results & Discussion
- References

Raw sequence data representing the majority of a bacterial genome can be obtained at a tiny fraction of the cost of a completed sequence. To demonstrate the utility of such a resource, 870 single-stranded M13 clones were sequenced from a shotgun library of the *Salmonella typhi* Ty2 genome. The sequence reads averaged over 400 bases and sampled the genome with an average spacing of once every 5,000 bases. A total of 339,243 bases of unique sequence was generated (approximately 7% representation). The sample of 870 sequences was compared to the complete *Escherichia coli* K-12 genome and to the rest of the GenBank database, which can also be considered a collection of sampled sequences. Despite the incomplete *S. typhi* data set, interesting categories could easily be discerned. Sixteen percent of the sequences determined from *S. typhi* had close homologs among known *Salmonella* sequences ($P < 1e^{-40}$ in BlastX or BlastN), reflecting the proportion of these genomes that have been sequenced previously; 277 sequences (32%) had no apparent orthologs in the complete *E. coli* K-12 genome ($P > 1e^{-20}$), of which 155 sequences (18%) had no close similarities to any sequence in the database ($P > 1e^{-5}$). Eight of the 277 sequences had similarities to genes in other strains of *E. coli* or plasmids, and six sequences showed evidence of novel phage lysogens or sequence remnants of phage integrations, including a member of the lambda family ($P < 1e^{-15}$). Twenty-three sample sequences had a significantly closer similarity a sequence in the database from organisms other than the *E. coli*/*Salmonella* clade (which includes *Shigella* and *Citrobacter*). These sequences are new

candidate lateral transfer events to the *S. typhi* lineage or deletions on the *E. coli* K-12 lineage. Eleven putative junctions of insertion deletion events greater than 100 bp were observed in the sample, indicating that well over 150 such events may distinguish *S. typhi* from *E. coli* K-12. The need for automatic methods to more effectively exploit sample sequences is discussed.

INTRODUCTION

The complete sequencing of bacterial genomes has revolutionized microbiology. However, the current high cost of completely sequencing genomes has limited its application to important pathogens and commercially important bacteria. The majority of this cost is incurred because of the labor-intensive methods which must still be used to close gaps covering the last few percent of the genome and to reduce the error rate to below 0.1%.

In contrast, a partial sequence of a bacterial genome can be obtained at low cost (39). Our costs of sequencing indicate that a random sample of sequences equivalent to the size of a genome (1× coverage) can be obtained at 1 to 2% of the cost of complete sequencing of the genome. Such a 1× "sample sequence" captures approximately 63% of the genome in at least one strand, the average contig size is about 690 bases, and the average gap size about 400 bases (using equations in reference 30). The number of clones required for sample sequencing of a 1× genome equivalent is directly related to genome size. For example, the genome of *Salmonella typhi*, which is 4.78 Mbp (34), would require about 12,000 reads of 400 bases. Similarly, a 2× sample sequence costs about 2 to 4% of a completed sequence and represents about 86% of the genome. The average contig size is about 1,280 bases, and the average gap size is 200 bases. In a bacterium, this level of sampling would ensure that almost every cistron was represented among the sample sequences.

The low cost of partial coverage of genomes makes it possible to consider sample sequencing of multiple genomes within a species, genus, or family. When a completely sequenced genome and a closely related sample-sequenced genome are compared, it is possible to identify sequences in the sampled genome that are absent in the completely sequenced genome. In bacteria, evolutionary mechanisms include the lateral transfer of cistrons and other units many kilobases in length, sometimes from distant species or phage. Thus, the presence of entire cistrons in one genome that are absent in a related genome is a quite common occurrence in bacteria, and these differences often contribute to the differences in life strategies of related species (18, 31). If multiple loci are available from multiple related species, then it is also possible to identify some of the loci that appear to have a phylogeny different from that of the rest of the genome. These are potential lateral transfers of genes or cistrons to the lineage of one genome or deletion events in the completely sequenced genome that have occurred since they diverged from their common ancestor. The vast GenBank database can be considered a huge collection of sample sequences for these purposes.

Here, we chose *S. typhi* for a pilot sample-sequencing effort because its genome is closely related to a completely sequenced genome, namely, that of *Escherichia coli* K-12 (6), and because it is closely related to the partially completed sequence of *Salmonella typhimurium* (45). The majority of the *S. typhi*

and *E. coli* genomes are probably related by descent from their common ancestor, and these regions share an average of about 85% identity at the nucleotide level and are even more conserved at the amino acid level (50). Previous studies used discrepancies in the alignments of the genetic maps of *Salmonella* and *E. coli* or DNA-DNA hybridization between these genomes to estimate that anywhere from 20 to 50% of these genomes may not be related by descent from their common ancestor (11, 26, 43). Indeed, even within the *Salmonella enterica* group (which includes *S. typhi*), up to 20% of the genome has been estimated to consist of genes that are not shared between pairs of strains (29).

S. typhi is of particular interest because it causes typhoid fever, a severe and sometimes fatal disease in humans. The only known effective host is humans, and so traditional methods for studying virulence mutations in model hosts are not adequate. Thus, sample sequencing of *S. typhi* could be particularly illuminating because it can identify candidates for genes involved in virulence.

MATERIALS AND METHODS

Top
Abstract
Introduction
Materials & Methods
Results & Discussion
References

Cloning and sequencing. Five micrograms of genomic DNA was sonicated, end repaired, fractionated, and subcloned in M13 as described previously (56); 1,059 subclones were purified and sequenced by a fluorescence-based sequencing method. Sequencing used standard shotgun library production, automatic plaque picking and DNA preparation, and short reads on an ABI 377 DNA sequencer. The cost was estimated at $1.89 per sequence read ($1.10 for supplies and $0.79 for labor), yielding a total cost of $2,000 for the sequence production. The success rate (percentage of subclones that provided high-quality sequence) for this library was 82.1%. The resulting raw sequence reads were processed by the program Automated Sequence Processor to remove low-quality traces and the X-Windows version of the Genome Assembly Program to assemble any overlapping reads. The 6% redundancy observed for the library is expected at this level of sampling (approximately 0.07-fold) (30). These shotgun sequencing methods are more fully described elsewhere (56).

Comparison to the GenBank database. At present there is a lack of good tools to pick out the most interesting from among a large number of sample sequences by using comparison with completed genomes and with other sequences in the database. Thus, we adapted data from the most readily available tools, the Blast suite of programs (reference 3 and references therein).

Each sample sequence was compared to the complete genome sequence of *E. coli* K-12 (6) by using BlastN 1.0 and TBlastX 1.0 and to the *E. coli* K-12 open reading frame (ORF) sequences by using BlastX 2.0. In addition, each sequence was compared to the entire GenBank nucleotide and amino acid databases with BlastN and BlastX, respectively, using the Blast server at the Genome Sequencing Center, Washington University, St. Louis, Mo. These data are available at http://genome.wustl.edu/gsc/bacterial/Salmonella.html. The data were further processed to show only the most significant match for each sequence, using Microsoft Word 6.0 with the assistance of macros. In some cases, matches with previously sequenced *Salmonella* sequences were removed first. The best hits for each search were entered into an Excel spreadsheet.

Significance thresholds for putative orthologs in *E. coli* K-12 and putative paralogous comparisons. The 870 sampled sequences were ranked by the significance score of their match with the *E. coli* K-12 genome. Putative orthologs were defined empirically as those matches that achieved $P < 1e^{-40}$, using either BlastN, BlastX, or TBlastX. A similar process was used to determine the number of orthologs with *Salmonella* sequences in the database. The threshold chosen was based on the fact that with sequence reads of 200 to 500 bases, this significance score always translated to a homology of greater than 60% nucleotide or amino acid identity spanning 200 or more bases, which is within the range expected for orthologous comparisons (50).

Alignments that yielded scores of $P > 1e^{-20}$ in all the three Blast search methods generally represented less than 60% nucleotide identity over a span of about 200 bases or less than 60% amino acid identity in a span of 60 amino acids (or a lower similarity in a longer alignment). When scores of $P > 1e^{-20}$ occurred in an *S. typhi* versus *E. coli* K-12 comparison, these were classified as putative paralogous comparisons.

Detecting potential lateral transfer and deletion events. For each of the 870 sample sequences, the ratios of the most significant score in *E. coli* K-12 and the most significant score in the rest of the GenBank database (other than *Salmonella*) were determined and ranked. These ratios were calculated from both BlastN and BlastX scores. The best examples of potential lateral transfer or deletion events were identified by first considering only those sample sequences (i) that had a significance of match of $P < 1e^{-15}$ in either BlastN or BlastX with an organism other than *E. coli* K-12 (or *Salmonella*) and (ii) where the match with this other organism had at least a 10,000-fold greater significance score than the best match in *E. coli* K-12, using both BlastN and BlastX. Amino acid similarities in the text are reported as single ratios that accumulate all nonoverlapping patches of similarity detected by the BlastX program.

RESULTS AND DISCUSSION

Top
Abstract
Introduction
Materials & Methods
Results & Discussion
References

As a preliminary demonstration of the utility of bacterial sample sequence resources, we sequenced 1,059 M13 clones from *S. typhi* Ty2. There were 870 reads of acceptable quality. The average read length was over 400 bases. These 870 clones melded into 791 contigs of 339,243 bp, representing about 7% of the genome.

The sequences were searched against the entire GenBank database, including the completed *E. coli* K-12 genome, using BlastX and BlastN. We found a continuum of similarities, ranging from a high degree of homology to no significant similarity, reflecting different evolutionary origins or different rates of divergence of the sequences.

Of the 870 sample sequences, a total of 135 (16%) had a presumed ortholog among sequences in various *Salmonella* serovars that were already in the database ($P < 1e^{-40}$) (Table 1). Sequences with these Blast scores reflect the cumulative proportion of genomes from various *Salmonella* serovars that had previously been sequenced in targeted projects.

View this table:
[in this window]
[in a new window]

TABLE 1. Similarities of *S. typhi* sample sequences to sequences in the public databases

There were 411 sequences (47%) that had highly significant homologies with the complete sequence of the *E. coli* K-12 genome ($P < 1e^{-40}$). These are presumably orthologs that diverged from a common ancestor of *E. coli* and *S. typhi*, although it is also possible that a few of these sequences are lateral transfer events between these two lineages since their divergence. The latter events would be characterized by exceptionally high conservation of DNA sequence compared to that found for typical orthologs, which are about 15% divergent in DNA sequence.

A total of 593 (68%) of the sample sequences had homologies with *E. coli* K-12 that were more significant than $P < 1e^{-20}$ (Table 1). Thus, 227 sequences (32%) had a less significant homology with *E. coli* K-12. At a threshold of $P > 1e^{-20}$, the match between sequences is sufficiently poor that they may not reflect true orthologs (by descent) between *Salmonella* and *E. coli* even after considering random fixation of mutations and errors in the sample sequences. Thus, 32% is perhaps an underestimate of the total amount of sequences in these genomes that are not homologous by descent from a common ancestor.

Most or all of these 227 sample sequences are presumably from "loops" in the *S. typhi* genome that distinguish this genome from the *E. coli* K-12 genome. Some of these 227 sequences may have been acquired by *S. typhi* since the divergence from the common ancestor with *E. coli*, while others may have been preserved in at least part of the *Salmonella* lineage but deleted in at least the K-12 part of the *E. coli* lineage. Among these 227 sample sequences, there are at least 10 examples that matched sequences in known loops from *Salmonella* that had already been characterized by other researchers. For example, hb59d06.s1 and hb59d10.s1 are almost identical to parts of *rfbG* (CDP-glucose 4,6-dehydratase gene) from the O-antigen cluster of *S. typhimurium* (22). hb59h10.s1 and hb60e06.s1 are almost identical to the *ssaR* gene from the type III secretion system apparatus of *S. typhimurium*. This is part of pathogenicity island 2 and is not present in *E. coli* K-12 or *Salmonella bongori*, which diverged at the first branch point in the *Salmonella* lineage. Pathogenicity island 2 was probably acquired after this divergence by horizontal transfer from an unknown source (19).

One interesting subset of this class of nonorthologs is the sequences that have no apparent homologs in the entire database. There were 155 sequences (18%) that had similarities less significant than $1e^{-5}$, a level at which the significance of any alignments are unreliable. These entirely novel sequences of no known function which occur in *S. typhi* but not *E. coli* K-12 presumably include some genes encoding novel functions.

Three-way comparisons. Pairwise comparisons of Blast significance scores are far from a foolproof strategy to detect potential lateral transfer and deletion events. Although most known genes in the *Salmonella* and *E. coli* genomes are closely homologous and these shared genes average about 85% identity at the nucleotide level, the random fixation of mutations means that genes that are related by

descent from the common ancestor vary widely around the mean of 85% identity. Lateral transfer is an ongoing process in these species and can occur between *E. coli* and *Salmonella*, between one of these species and other closely related genomes (such as *Citrobacter* and *Shigella*), or between one of these species and a more distantly related genome. Thus, similarities between paralogous genes (including those due to lateral transfer) lie on a continuum that overlaps the similarities between genes that are related by descent from the common ancestor. As a consequence, estimates of the level of lateral transfer by using comparisons between sample sequences from *S. typhi* and the complete *E. coli* K-12 genome (or, in general, between any two genomes) are inherently unreliable. Another serious limitation of using Blast scores for the whole read length of each sample sequence to rank sample sequences is that the sample sequences have different read lengths and the significance scores are sensitive to the length of the homology detected. In the analyses discussed above, we have tried to avoid these problems by adjusting the significance thresholds to reflect these facts. However, these limitations can be more effectively mitigated by using a phylogenetic comparison with a third species.

The key to identifying novel paralogous comparisons is to have a third reference sequence from an outgroup species. In most cases, closely related sequences in two ingroup species will be more similar to each other than either is to any sequence in the outgroup species. However, if this is not true, then a potential lateral transfer or deletion event is revealed. It might be argued that the sample sequences are short and contain occasional errors, and so this might be an unreliable strategy. However, it should be noted that insertion/deletion errors in the sample sequence will be "private" (i.e., uninformative), and accidental matches of miscalled bases will occur with approximately equal frequency in each true homolog. Both types of error will not typically bias a match to one homolog in the database versus another. The best matches detected will typically reflect the closest similarities that would be seen if the sample sequence were error free, although the apparent genetic distance of the sampled sequence may be exaggerated by sequencing errors.

The best examples of potential lateral transfer or deletion events are discussed below. These examples were identified by stringent criteria in which the Blast score in *E. coli* was much less significant than the Blast score for some other sequence in the GenBank database (see Materials and Methods). The criteria undoubtedly removed some legitimate examples of potential lateral transfer events (or examples of deleted sequences in the K-12 lineage where the best score would be a paralogous comparison). Nevertheless, these criteria concentrated the search toward the best-supported examples.

Only new relationships that could not be deduced previously from the sequences already in the databases are discussed below and presented in Table 2. Thus, those sample sequences that were homologous to known *Salmonella* sequences were dropped from consideration.

View this table: [in this window] [in a new window]

TABLE 2. Comparison of *S. typhi* sample sequences with the public databases[a]

Sequences found in some *E. coli* strains but not found in strain K-12. Using the above criteria, we found eight *S. typhi* sample sequences that had better matches with sequences from *E. coli* strains other than *E. coli* K-12 and which did not occur in known *Salmonella* sequences (Table 2). Clones hb53h05.s1, hb56g12.s1, and hb57e10.s1 are similar to three enzymes in an aromatic degradative pathway of some *E. coli* strains where the cluster of genes occurs as an insertion relative to *E. coli* K-12 (41, 42, 44). This is presumably an example of how some *E. coli* strains, and apparently at least this one *Salmonella* strain, have become adapted to a new nutritional source by the recruitment of a catabolic cassette.

The strain of *S. typhi* that we used has no known plasmids. Nevertheless, there were four sample sequences that have their closest similarities to genes found on plasmids in *E. coli*. hb56b07.s1 has similarity to the immunity protein of the ColE7 plasmid (12) which is found in a few strains of *E. coli* (56/74 [76%] similarity to this 84-amino-acid peptide). hb57h01.s2 has similarity to the transfer operon gene, *traF*, of the conjugative F plasmid found in some *E. coli* strains (57). hb54b06.s1 has a patch of moderate similarity (47/71 [66%]) to a plasmid-associated chaperone gene of enteroaggregative *E. coli* (47). hb62d06.s1 has an ORF similar to *ipgD* of *Shigella sonnei* (cumulative 87/137 [64%] similarity), the gene for a secreted protein on a virulence plasmid proximal to *mxi* (2). hb55b09.s1 is highly similar to the transposon- and plasmid-borne citrate utilization gene, *citB*, found in *Klebsiella* spp. and in some *E. coli* strains other than K-12 (21). Another clone, hb53b07.s1, may be more closely related to *citB* in *Klebsiella* (93% similarity in 131 amino acids) than *E. coli* (73% similarity in 141 amino acids). Tricarboxylic acids are used in many *Salmonella* serovars but not in *E. coli*; citrate is used as a carbon source. The *tet* operon at 60 min is sequenced in *Salmonella*. The *tetII* genes and *citB* and *citA* map at 17 min but are not sequenced in *S. typhimurium* LT2; thus, these genes probably are present in *S. typhimurium* and in *Klebsiella* but missing from *E. coli*.

Probably the sequences noted above are integrated in the *S. typhi* genome rather than being on a previously unknown plasmid. Some of these sequences presumably represent further examples of sequences that are found on plasmids in one bacterium but in the genomes of other bacteria. Sequences recruited to the genome by integration of plasmids are probably a major source of the loops that distinguish bacterial genomes.

Sequences similar to phage. The *S. typhi* Ty2 genome does not have any previously known integrated prophage. Nevertheless, limited sequence similarity to various bacteriophages or retrons was found in this genome. hb58g10.s1 has some similarity to a retron-associated sequence (32) and to a bacteriophage P2 putative vertex protein (44/55 [80%] similarity) (33). This is the first example of a sequence that may be associated with a retron in *Salmonella*. hb55g10.s1 has some similarity to an *E. coli* retron methylase (92/144 [64%] similarity). It is possible that this is another retron-associated sequence in *S. typhi*. It is related to *dam* from *Serratia marcescens* (38) (cumulative 76/124 [60%] similarity), *S. typhimurium* (81/123 [66%] similarity), and *E. coli* (73/123 [59%] similarity) but is less related to these proteins than they are to each other (>80% similarity over 200 amino acids). The other end was selected for sequencing and has a patch of DNA sequence 143/149 (95%) identical to a sequence in the same *E. coli* retron and a conceptual translation which is 35/43 (81%) identical to phage P2 terminase ATPase subunit.

hb53e05.s1 has some similarity to bacteriophage P2 in the proposed endonuclease subunit of terminase (71/113 [63%] similarity) (33). hb55d08.s1 is related to the *gam* gene of bacteriophage Mu (36/55 [65%] identities). The *gam* gene encodes a protein which protects linear double-stranded DNA from exonuclease degradation in vitro and in vivo (1). hb60b03.s1 has similarity to a coliphage 186 putative tail fiber assembly protein (61/86 [71%] amino acid similarity) (58). Finally, hb57g10.s2 is about 60% identical to a tail protein of bacteriophage lambda at the DNA level and 60% identical at the protein level.

Despite these similarities to phage, it should be noted that the typical lysogenic phage is many tens of kilobases in length, and so any complete prophage in the genome should yield a number of sequences from a sample of the size that we used (one clone every 5 kb). Thus, complete genomes from close homologs of known bacteriophages are unlikely to occur in the Ty2 genome. The sequences observed may be remnant parts of an ancient prophage that the genome has preserved for its own needs.

Sequences similar to other enterobacteriaceae. Another set of sample sequences have their closest similarities in the database to sequences that have been characterized in enterobacteria outside the *E. coli Shigella Salmonella Citrobacter* clade. Such sequences are of particular interest because they may represent deletion events in the *E. coli* K-12 lineage or insertion events in the *S. typhi* lineage so that the gene in *Salmonella* shares an ancestor with an organism other than *E. coli*. Either way, their phylogeny may not be the same as the phylogeny of the majority of the genome shared by *Salmonella* and *Escherichia*. The similarities in the database can give clues as to the function of these sequences in *S. typhi*, a function that may not have an exact counterpart in the *E. coli* K-12 genome.

Among this class of sample sequences are a number that are most closely similar to genes in the close sibling genus, *Klebsiella*. hb57b09.s1 and hb91f10.s1 contain ORFs with patches of close similarity to *citA* (8), the sensor kinase, in *Klebsiella* (113/134 and 153/172 [84 and 89%] similarity, respectively) and much less similarity to the *E. coli* K-12 sensor kinase gene *citA* (51/93 and 111/162 [55 and 69%] similarity, respectively).

hb53d08.s1 and hb58d01.s1 contain different portions of an ORF that is closely related to the arylsulfate sulfotransferase of the phylogenetically very distant bacterium *Campylobacter jejuni* (108/135 and 132/152 [80 and 87%] amino acid similarity, respectively) (59). These ORFs are less related to the *Klebsiella* protein in the same region (<70% similarity) (4). hb57h08.s2 contains sequences that have similarity to the arylsulfate sulfotransferase protein of *Klebsiella* (66/95 [69%] similarity). hb58b02.s1 is almost identical to the disulfide isomerase of *Klebsiella* in the same arylsulfate metabolic complex.

To obtain further supporting evidence for a paralogous comparison, as part of our effort to sequence the entire genome of *S. typhimurium*, the sequencing of the region corresponding to the arylsulfatase in *E. coli* K-12 is in progress (unpublished data). This gene and an adjacent regulator protein are absent in the corresponding part of the *S. typhimurium* genome, further supporting the possibility that the *Salmonella* and *E. coli* genes are different in phylogenetic history and location in the genome.

hb58e07.s1 and hb62d05.s1 have similarity to *rthD*, a putative protein of unknown function (36/38 and

63/72 [95 and 88%] similarity, respectively) in the 6.6-kb *rfb* gene cluster from *Klebsiella pneumoniae* serotype O1 (*rfbKp*O1). This cluster contains six genes whose products are required for the biosynthesis of a lipopolysaccharide O antigen (14). hb62f04.s1 is closely related to another gene in the cluster, *rfbF*, encoding the galactosyltransferase protein of *K. pneumoniae* (133/156 [85%] similarity). This cluster would be expected to be missing from *E. coli* K-12, and many other enteric organisms, which do not put rhamnose into their lipopolysaccharide. An *rfb* cluster has been cloned from *S. typhimurium* (22), but the *Klebsiella rfbF* and *S. typhi* sequences are not related to this cluster.

Some rather unexpected similarities occur between *S. typhi* sample sequences and sequences in other enterobacteria or related proteobacteria. For example, hb55f03.s1 is remarkable in that it shares very significant similarity with indolepyruvate decarboxylase from *Enterobacter cloacae*, another enterobacterium (164/217 at the DNA level; cumulative patches of similarity of 111/124 [90%] at the amino acid level). This enzyme is used to convert indole-3-pyruvic acid to indole-3-acetic acid, a well-known plant hormone. Other enterobacteria that have this gene are *Enterobacter agglomerans* strains, *Pantoea agglomerans*, *Klebsiella aerogenes*, and *Klebsiella oxytoca* (61), some of which are opportunistic pathogens of humans.

hb91e02.s1 is related to the high-affinity outer membrane ferrioxamine receptor *foxA* of *Yersinia enterocolitica* (63/105 [60%] similarity) (5) and has slightly less significant similarity with the ferrichrome-iron receptor precursor of *E. coli* (57/99 [58%] similarity). Either this is a paralogous comparison between *E. coli* and *S. typhi* or the gene has been under strong selective pressure to diverge quickly in one or both of these species.

Sequences similar to distantly related organisms. hb59h11.s1 contains an ORF related to an accessory colonization factor of *Vibrio cholerae* (similarity, 78/118 [66%]) which is probably related to the methyl-accepting chemotaxis proteins (16). hb56f06.s1 is closely related to the histidine ammonia-lyase (*hutH*) gene of *Pseudomonas putida* (patches totaling 87/117 [74%] amino acid similarity) (15). hb56h08.s1 is similar to a mandelate racemase of *P. putida* (77/139 [55%] similarity) (37).

Three *S. typhi* sample sequences have their closest similarity in the database with species of the phylogenetically very distant bacterium *Haemophilus*. hb56b09.s1 shares some similarity with the *Haemophilus influenzae* transport ATPase protein *cydC* (51/84 [61%] similarity) and a weaker similarity with the *cydC* from *E. coli* (42/69 [61%] similarity). This latter protein is an ABC (ATP-binding cassette) family membrane transporter necessary for the formation of the cytochrome *bd* quinol oxidase (40). hb55h11.s1 has weak similarities to a leukotoxin secretion ATP-binding protein of *Haemophilus actinomycetemcomitans* (42/88 [48%] similarity) (28) and a much weaker similarity to an *E. coli* hemolysin secretion ATP-binding protein not found in the K-12 strain. These are part of cytolytic toxin complexes. hb58f11.s1 is similar to a *Haemophilus* hypothetical protein of unknown function (cumulative similarities, 82/127 [65%]).

hb53f11.s1 displays a remarkable similarity to the chitinase proteins of a number of phylogenetically very distant bacteria. The similarity to the chitinase of *Aeromonas caviae* (51) is 82/147 (56%) extending over almost the whole protein. It is hard to imagine what the purpose of the related gene might be in *S*

typhi. Perhaps the gene in *S. typhi* has a different substrate. The other end of this clone was selected for sequencing and proved to be homologous to *pepA* at 21 min.

The closest known similarities for a number of other sample sequences occur in other phylogenetically very distant bacteria. hb57b04.s1 is related to a protein in the *trpA* region of *Buchnera aphidicola* (88/130 [68%] similarity) (27). The function of that protein is not known, but it is related to the *yedA* (*E. coli*) and *yexC* (*Bacillus subtilis*) genes, which are thought to encode integral membrane proteins. hb62g03.s1 shows some similarity to the *B. subtilis* choline transport system ATPase (*opuBA*) and *proV* (60/78 [77%] similarity) (13) but little similarity to genes in *E. coli* or *Salmonella*, though there is weak similarity to a hypothetical ABC transporter (*ychV*) in *E. coli* (51/75 [68%] similarity). hb62h09.s1 has regions of similarity with an ORF involved in conjugal transfer in the *oriT* region of a streptococcal plasmid (52/94 [55%] similarity) (53). Finally, hb56e11.s1 and hb55a12.s1 overlap and share similarity to the voltage-dependent potassium channel alpha subunit from many eukaryotes. The best of these similarities is to *Caenorhabditis elegans*, where patches of 66/102 (65%) similarity are observed. This sequence is weakly homologous (33/94 [35%] similarity) to a homolog of eukaryotic potassium channel proteins previously noted in *E. coli* (36). The other end of this clone was sequenced and found to encode cardiolipin synthetase, which maps at 28 min in *E. coli* K-12.

Candidate insertion/deletion junctions. Some sample sequences will consist of junctions of insertion/deletion events that distinguish *S. typhi* from *E. coli* K-12 or from other *Salmonella* strains. At present, the tools to conveniently find candidate junction sequences are under construction (36a). However, we noted seven examples by visual inspection of BlastN alignments of the *S. typhi* clones with the *E. coli* K-12 genome (Table 3). Some of these insertion deletion events may only be 100 bases in length, but some may be the junctions of very large insertions in *S. typhi*.

TABLE 3. Putative insertion/deletion junction fragments

View this table:
[in this window]
[in a new window]

Four other clones in which the sequence read from one end had interesting homologies in the database and poorer homology with *E. coli* K-12 were chosen for sequencing from the other end of the clone. These were hb53f11, containing a chitinase homolog, hb55f03, containing an indolepyruvate decarboxylase homolog, hb55g10, containing a retron-associated sequence, and hb56e11, containing a voltage-dependent potassium channel alpha subunit homolog (Table 2). Surprisingly, in all four cases the other end of the clone was closely homologous to a known *E. coli* sequence, indicating the location of the junction between unique and shared sequences within the ca. 1.5-kb clone. If these unique regions in *S. typhi* that are not present in *E. coli* were generally many kilobases in length, then clones that contained unique sequences at one end would generally also contain unique sequences at the other end. The fact that all four clones contained junctions between unique and shared DNA suggests that many of the genes that distinguish *E. coli* and *S. typhi* may be found as single genes or in small groups of a few genes. Thus, the cloning of a number of large pathogenicity islands of 10 kb or more, each containing many

genes that distinguish *Salmonella* from *E. coli*, may lead to an exaggerated impression of the average number of genes in each insertion deletion event between these species. Indeed, we found that 7% of the genome contains at least 11 insertion deletion junctions for unique sequences over 100 bp. By extrapolation, there should be at least 157 such events that distinguish *S. typhi* from *E. coli* K-12.

One clone appears to span a junction of a region that differs between *S. typhimurium* and *S. typhi*. One portion of hb58f04.s1 encodes the *S. typhimurium* phosphoglycerate transport system activator (*pgtA*) gene (60), whereas another portion shows 38/51 (75%) similarity to a tail spike protein from bacteriophage P22 (52).

To confirm such junctions, the portion that is apparently unique to *S. typhi* can be used as a probe in a Southern blot of *S. typhimurium* or *E. coli* DNA to determine if it is absent in these genomes. Alternatively, if the insertion in *S. typhi* is less than 10 kb, then PCR primers that are a few hundred bases apart in *S. typhimurium* or *E. coli* should yield a much larger PCR product spanning the insertion in *S. typhi*.

Improvements in the search strategy. Many examples of sample sequences that were more closely related to sequences in the database other than that of *E. coli* K-12 were undoubtedly missed by the methods used here. For example, while the comparison with *E. coli* K-12 is with a complete genome so the sample sequence will generally align over the maximum possible region of homology, the rest of the database is fragmentary. Each time a *Salmonella* sequence overlaps only partly at one end of another sequence in the database, the Blast significance will be lower even though the region of match is excellent. This limitation would be circumvented by a program that could compare sample sequences to each other and to the fragmentary data in the GenBank database by using only regions of similarity shared by all three (or more) sequences rather than the pairwise comparisons used here.

The comparison of sample sequences from multiple organisms to one or more closely related completed genomes could be an effective strategy for discovering genes that distinguish species. As analytical tools are improved, it should be possible to ask even more sophisticated questions with sample sequence data. For example, for pathogenic bacteria, one of the most interesting applications would be to compare the rates of evolution of loci across the genome. Cell surface proteins in pathogenic bacteria are exposed to the immune system of the host. This can lead to a selective pressure that is greater than that experienced by most other genes in the genome (9). Analytical tools that could align multiple sequences and then compare the rates of evolution of synonymous and nonsynonymous codons only in the regions shared by all of the sample sequences would allow detection of candidate loci that may have undergone accelerated evolution. Where these loci exist they would be of particular interest for further study as potentially vital parts of the pathogenic mechanism and as immunologic targets.

Pathways and structures. Databases of the known metabolic pathways in bacteria and homologs for genes in these pathways have been assembled (23, 24, 48, 49). As these databases grow to encompass all functions of the cell, one potential use of sample sequences is to determine whether metabolic or transport pathways, signal transduction pathways, or particular physical structures are present in a bacterium.

With these resources, what is the amount of sample sequencing needed to determine whether a pathway or structure is present in a bacterium? Certainly, less than the complete genome is necessary because one need sample only one or a few genes in a pathway or structure to be confident that the pathway or structure is present. Furthermore, one does not need to determine the complete sequence of an ORF if it is a close homolog of a known gene in another bacterium. Perhaps one need sample a stretch of only 50 amino acids (of 98% accurate sequence) from a gene to be able to assign those that have a close homolog already in the database. As our knowledge of pathways and structures grows, it should be possible to determine the probable presence (or probable absence) of an increasing number even with only 50% of the genome represented in a sample sequence. Furthermore, as more sample sequences are obtained, it should be easier to assign homologs. Such a sample sequence can be obtained for as little as $25,000 at a sequencing center, and the price can be expected to continue to fall.

It is interesting that because genes tend to be clustered in cistrons, a shotgun sample sequence of 50% of a genome is better for these purposes than a complete sequence of one half of the genome, as well as being much less expensive. The latter strategy will miss entire cistrons in the half of the genome that has not been sequenced. In contrast, the shotgun approach will sample a small portion of virtually all cistrons.

In conclusion, although there are certain limitations of sample sequences, these limitations are more than counterbalanced by the knowledge that can be gained at very low cost. It is hoped that sample sequencing will begin in earnest and that the bioinformatics needed to fully exploit sample sequences will be developed.

ACKNOWLEDGMENTS

M.M. was partially supported by a generous gift from Sidney Kimmel and by NIH grants CA-68822, NS-33377, AI43283, and AI-34829.

We thank Ken Sanderson for many important discussions and for critically reviewing the manuscript. We thank Michael Nhan for generating the web page and for the batch searches, and we thank other members of the Genome Sequencing Center for technical assistance.

FOOTNOTES

* Corresponding author. Mailing address: Sidney Kimmel Cancer Center, 10835 Altman Row, San Diego, CA 92121. Phone: (619) 450-5990, ext. 280. Fax: (619) 550-3998. E-mail: mmcclelland *at* skcc.org.

Editor: J. G. Cannon

REFERENCES

Top
Abstract
Introduction
Materials & Methods
Results & Discussion
References

1. **Akroyd, J. E., E. Clayson, and N. P. Higgins.** 1986. Purification of the gam gene-product of bacteriophage Mu and determination of the nucleotide sequence of the gam gene. Nucleic Acids Res. **14:**6901-6914[Abstract].

2. **Allaoui, A., R. Menard, P. J. Sansonetti, and C. Parsot.** 1993. Characterization of the *Shigella flexneri ipgD* and *ipgF* genes, which are located in the proximal part of the *mxi* locus. Infect. Immun. **61:**1707-1714[Abstract].

3 **Altschul, S. F., T. L. Madden, A. A. Schaffer, et al.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:**3389-3402[Abstract Full Text].

4 **Baek, M. C., S. K. Kim, D. H. Kim, B. K. Kim, and E. C. Choi.** 1996. Cloning and sequencing of the *Klebsiella* K-36 astA gene, encoding an arylsulfate sulfotransferase. Microbiol. Immunol. **40:**531-537[Medline].

5 **Baumler, A. J., and K. Hantke.** 1992. Ferrioxamine uptake in *Yersinia enterocolitica*: characterization of the receptor protein FoxA. Mol. Microbiol. **6:**1309-1321[Medline].

6. **Blattner, F. R., G. Plunkett, C. A. Bloch, et al.** 1997. The complete genome sequence of *Escherichia coli* K-12. Science **277:**1453-1474[Abstract Full Text].

7. **Bott, M., and P. Dimroth.** 1994. *Klebsiella pneumoniae* genes for citrate lyase and citrate lyase ligase: localization, sequencing, and expression. Mol. Microbiol. **14:**347-356[Medline].

8 **Bott, M., M. Meyer, and P. Dimroth.** 1995. Regulation of anaerobic citrate metabolism in *Klebsiella pneumoniae*. Mol. Microbiol. **18:**533-546[Medline].

9. **Boyd, E. F., J. Li, H. Ochman, and R. K. Selander.** 1997. Comparative genetics of the *inv-spa* invasion gene complex of *Salmonella enterica*. J. Bacteriol. **179:**1985-1991[Abstract].

10. **Brandl, M. T., and S. E. Lindow.** 1996. Cloning and characterization of a locus encoding an indolepyruvate decarboxylase involved in indole-3-acetic acid synthesis in *Erwinia herbicola*. Appl. Environ. Microbiol. **62:**4121-4128[Abstract].

11. **Brenner, D. J.** 1984. Enterobacteriaceae, p. 408-420. *In* N. R. Krieg, and J. G. Holt (ed.), Bergey's manual of systematic bacteriology. Williams & Wilkins, Baltimore, Md.

12. **Chak, K. F., W. S. Kuo, F. M. Lu, and R. James.** 1991. Cloning and characterization of the ColE7 plasmid. J. Gen. Microbiol. **137:**91-100[Medline].

13. **Chung, Y. J., and J. N. Hansen.** 1992. Determination of the sequence of *spaK* and identification of a promoter in the subtilin (*spa*) operon in *Bacillus subtilis*. J. Bacteriol. **174:**6699-6702[Abstract].

14. **Clarke, B. R., D. Bronner, W. J. Keenleyside, W. B. Severn, J. C. Richards, and C. Whitfield.** 1995. Role of Rfe and RfbF in the initiation of biosynthesis of D-galactan I, the lipopolysaccharide O antigen from *Klebsiella pneumoniae* serotype O1. J. Bacteriol. **177:**5411-5418[Abstract].

15. **Consevage, M. W., and A. T. Phillips.** 1990. Sequence analysis of the *hutI* gene encoding histidine ammonia-lyase in *Pseudomonas putida*. J. Bacteriol. **172:**2224-2229[Medline].

16. **Everiss, K. D., K. J. Hughes, M. E. Kovach, and K. M. Peterson.** 1994. The *Vibrio cholerae actB* colonization determinant encodes an inner membrane protein that is related to a family of

signal-transducing proteins. Infect. Immun. **62:**3289-3298[Abstract].

17. **Fleischmann, R. D., M. D. Adams, O. White, et al.** 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science **269:**496-512[Medline].
18. **Hacker, J., G. Blum-Oehler, I. Muhldorfer, and H. Tschape.** 1997. Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. Mol. Microbiol. **23:**1089-1097[Medline].
19. **Hensel, M., J. E. Shea, A. J. Baumler, C. Gleeson, F. Blattner, and D. W. Holden.** 1997. Analysis of the boundaries of *Salmonella* pathogenicity island 2 and the corresponding chromosomal region of *Escherichia coli* K-12. J. Bacteriol. **179:**1105-1111[Abstract].
20. **Hsu, M. Y., M. Inouye, and S. Inouye.** 1990. Retron for the 67-base multicopy single-stranded DNA from *Escherichia coli*: a potential transposable element encoding both reverse transcriptase and Dam methylase functions. Proc. Natl. Acad. Sci. USA **87:**9454-9458[Abstract].
21. **Ishiguro, N., M. Sasatsu, T. K. Misra, and S. Silver.** 1988. Promoters and transcription of the plasmid-mediated citrate-utilization system in *Escherichia coli*. Gene **68:**181-192[Medline].
22. **Jiang, X. M., B. Neal, F. Santiago, S. J. Lee, L. K. Romana, and P. R. Reeves.** 1991. Structure and sequence of the rfb (O antigen) gene cluster of *Salmonella* serovar *typhimurium* (strain LT2). Mol. Microbiol. **5:**695-713[Medline].
23. **Karp, P. D.** 1998. Metabolic databases. Trends Biochem. Sci. **23:**114-116[Medline].
24. **Karp, P. D., M. Riley, S. M. Paley, A. Pellegrini-Toole, and M. Krummenacker.** 1998. EcoCyc: encyclopedia of *Escherichia coli* genes and metabolism. Nucleic Acids Res. **26:**50-53[Abstract/Full Text].
25. **Koga, J., T. Adachi, and H. Hidaka.** 1991. Molecular cloning of the gene for indolepyruvate decarboxylase from *Enterobacter cloacae*. Mol. Gen. Genet. **226:**10-16[Medline].
26. **Krawiec, S., and M. Riley.** 1990. Organization of the bacterial chromosome. Microbiol. Rev. **54:**502-539[Medline].
27. **Lai, C. Y., P. Baumann, and N. A. Moran.** 1995. Genetics of the tryptophan biosynthetic pathway of the prokaryotic endosymbiont (*Buchnera*) of the aphid *Schlechtendalia chinesis*. Insect Mol. Biol. **4:**47-59[Medline].
28. **Lally, E. T., E. E. Golub, I. R. Kieba, et al.** 1991. Structure and function of the B and D genes of the *Actinobacillus actinomycetemcomitans* leukotoxin complex. Microb. Pathog. **11:**111-121[Medline].
29. **Lan, R., and P. R. Reeves.** 1996. Gene transfer is a major factor in bacterial evolution. Mol. Biol. Evol. **13:**47-55[Abstract].
30. **Lander, E. S., and M. S. Waterman.** 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics **2:**231-239[Medline].
31. **Lawrence, J. G., and J. R. Roth.** 1996. Evolution of coenzyme B12 synthesis among enteric bacteria: evidence for loss and reacquisition of a multigene complex. Genetics **142:**11-24[Abstract].
32. **Lim, D.** 1995. Analysis of a retron EC86 and EC67 insertion site in *Escherichia coli*. Plasmid **34:**58-61[Medline].
33. **Linderoth, N. A., R. Ziermann, E. Haggard-Ljungquist, G. E. Christie, and R. Calendar.** 1991. Nucleotide sequence of the DNA packaging and capsid synthesis genes of bacteriophage P2. Nucleic Acids Res. **19:**7207-7214[Medline].
34. **Liu, S. L., and K. E. Sanderson.** 1995. Rearrangements in the genome of the bacterium *Salmonella typhi*. Proc. Natl. Acad. Sci. USA **92:**1018-1022[Abstract].

35. **McCaman, M. T., and J. D. Gabe.** 1986. Sequence of the promoter and 5' coding region of pepN, and the amino-terminus of peptidase N from *Escherichia coli* K-12. Mol. Gen. Genet. **204:**148-152[Medline].

36. **Milkman, R.** 1994. An *Escherichia coli* homologue of eukaryotic potassium channel proteins. Proc. Natl. Acad. Sci. USA **91:**3510-3514[Abstract].

36a. **Miller, W.** Personal communication.

37. **Neidhart, D. J., G. L. Kenyon, J. A. Gerlt, and G. A. Petsko.** 1990. Mandelate racemase and muconate lactonizing enzyme are mechanistically distinct and structurally homologous. Nature **347:**692-694[Medline].

38. **Ostendorf, T., P. Cherepanov, M. Jekel, J. de Vries, and W. Wackernagel.** 1994. S. marcescens Sr 41 urf, dam and dod genes. GenBank accession no. X78412.

39 **Peterson, S. N., P. C. Hu, K. F. Bott, and C. A. Hutchison.** 1993. A survey of the *Mycoplasma genitalium* genome by using random sequencing. J. Bacteriol. **175:**7918-7930[Abstract].

40. **Poole, R. K., L. Hatch, M. W. Cleeter, F. Gibson, G. B. Cox, and G. Wu.** 1993. Cytochrome bd biosynthesis in *Escherichia coli*: the sequences of the *cydC* and *cydD* genes suggest that they encode the components of an ABC membrane transporter. Mol. Microbiol. **10:**421-430[Medline].

41 **Prieto, M. A., E. Diaz, and J. L. Garcia.** 1996. Molecular characterization of the 4-hydroxyphenylacetate catabolic pathway of *Escherichia coli* W: engineering a mobile aromatic degradative cluster. J. Bacteriol. **178:**111-120[Abstract].

42. **Prieto, M. A., and J. L. Garcia.** 1994. Molecular characterization of 4-hydroxyphenylacetate 3-hydroxylase of *Escherichia coli*. A two-protein component enzyme. J. Biol. Chem. **269:**22823-22829[Abstract].

43. **Riley, M., and K. E. Sanderson.** 1990. Comparative genetics of *Escherichia coli* and *Salmonella typhimurium*, p. 85-95. *In* K. Drlica, and M. Riley (ed.), The bacterial chromosome. American Society for Microbiology, Washington, D.C.

44. **Roper, D. I., and R. A. Cooper.** 1990. Subcloning and nucleotide sequence of the 3,4-dihydroxyphenylacetate (homoprotocatechuate) 2,3-dioxygenase gene from *Escherichia coli* C. FEBS Lett. **275:**53-57[Medline].

45. **Sanderson, K. E., S.-L. Liu, A. Hessel, and K. E. Rudd.** 1996. The genetic map of *Salmonella typhimurium*, edition VIII, p. 1903-1999. *In* F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella*: cellular and molecular biology, 2nd ed. ASM Press, Washington, D.C.

46. **Sanger, F., A. R. Coulson, G. F. Hong, D. F. Hill, and G. B. Petersen.** 1982. Nucleotide sequence of bacteriophage lambda DNA. J. Mol. Biol. **162:**729-773[Medline].

47. **Savarino, S. J., P. Fox, Y. Deng, and J. P. Nataro.** 1994. Identification and characterization of a gene cluster mediating enteroaggregative *Escherichia coli* aggregative adherence fimbria I biogenesis. J. Bacteriol. **176:**4949-4957[Abstract].

48. **Selkov, E., M. Galimova, I. Goryanin, et al.** 1997. The metabolic pathway collection: an update. Nucleic Acids Res. **25:**37-38[Abstract Full Text].

49. **Selkov, E., N. Maltsev, G. J. Olsen, R. Overbeek, and W. B. Whitman.** 1997. A reconstruction of the metabolism of *Methanococcus jannaschii* from sequence data. Gene **197:**GC11-GC26[Medline].

50. **Sharp, P. M.** 1991. Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: codon usage, map position, and concerted evolution. J. Mol. Evol. **33:**23-33[Medline].

51. **Sitrit, Y., C. E. Vorgias, I. Chet, and A. B. Oppenheim.** 1995. Cloning and primary structure of the *chiA* gene from *Aeromonas caviae*. J. Bacteriol. **177:**4187-4189[Abstract].

52. **Steinbacher, S., R. Seckler, S. Miller, B. Steipe, R. Huber, and P. Reinemer.** 1994. Crystal structure of P22 tailspike protein: interdigitated subunits in a thermostable trimer. Science **265:**383-386[Medline].

53. **Wang, A., and F. L. Macrina.** 1995. Characterization of six linked open reading frames necessary for pIP501-mediated conjugation. Plasmid **34:**206-210[Medline].

54. **Wang, A., and F. L. Macrina.** 1995. Streptococcal plasmid pIP501 has a functional *oriT* site. J. Bacteriol. **177:**4199-4206[Abstract].

55. **Wilson, R., R. Ainscough, K. Anderson, et al.** 1994. 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans*. Nature **368:**32-38[Medline].

56. **Wilson, R. K., and E. R. Mardis.** 1997. Shotgun sequencing. p. 396-454. *In* B. Birren, E. Green, S. Klapholz, R. Myers, and J. Roskams (ed.), Genome analysis: a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

57. **Wu, J. H., P. Kathir, and K. Ippen-Ihler.** 1988. The product of the F plasmid transfer operon gene, *traF*, is a periplasmic protein. J. Bacteriol. **170:**3633-3639[Medline].

58. **Xue, Q., and J. B. Egan.** 1995. DNA sequence of tail fiber genes of coliphage 186 and evidence for a common ancestor shared by dsDNA phage fiber genes. Virology **212:**128-133[Medline].

59. **Yao, R., and P. Guerry.** 1996. Molecular cloning and site-specific mutagenesis of a gene involved in arylsulfatase production in *Campylobacter jejuni*. J. Bacteriol. **178:**3335-3338[Abstract].

60. **Yu, G. Q., and J. S. Hong.** 1986. Identification and nucleotide sequence of the activator gene of the externally induced phosphoglycerate transport system of *Salmonella typhimurium*. Gene **45:**51-57[Medline].

61. **Zimmer, W., B. Hundeshagen, and E. Niederau.** 1994. Demonstration of the indolepyruvate decarboxylase gene homologue in different auxin-producing species of the Enterobacteriaceae. Can. J. Microbiol. **40:**1072-1076[Medline].

Infection and Immunity, September 1998, p. 4305-4312, Vol. 66, No. 9
0019-9567/98/$04.00+0

This article has been cited by other articles:

- Ffrench-Constant, R. H., Waterfield, N., Burland, V., Perna, N. T., Daborn, P. J., Bowen, D., Blattner, F. R. (2000). A Genomic Sample Sequence of the Entomopathogenic Bacterium Photorhabdus luminescens W14: Potential Implications for Virulence. *Appl. Environ. Microbiol.* 66: 3310-3329 [Abstract] [HTML]
- Schwartz, S., Zhang, Z., Frazer, K. A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., Miller, W. (2000). PipMaker---A Web Server for Aligning Two Genomic DNA Sequences. *Genome Res.* 10: 577-586 [Abstract] [HTML]
- Morrow, B. J., Graham, J. E., Curtiss, R. III (1999). Genomic Subtractive Hybridization and

Abstract of this Article
Reprint (PDF) Version of this Article
Similar articles found in:
IAI Online
PubMed
PubMed Citation
This Article has been cited by:
Search Medline for articles by:
McClelland, M. Wilson, R. K.
Alert me when
new articles cite this article
Download to Citation Manager

Selective Capture of Transcribed Sequences Identify a Novel Salmonella typhimurium Fimbrial Operon and Putative Transcriptional Regulator That Are Absent from the Salmonella typhi Genome. *Infect. Immun.* 67: 5106-5116 [Abstract] [HTML]

- Benson, N. R., Wong, R. M.-Y., McClelland, M. (2000). Analysis of the SOS Response in Salmonella enterica Serovar Typhimurium Using RNA Fingerprinting by Arbitrarily Primed PCR. *J. Bacteriol.* 182: 3490-3497 [Abstract] [HTML]
- Ibanez-Ruiz, M., Robbe-Saule, V., Hermant, D., Labrude, S., Norel, F. (2000). Identification of RpoS (sigma S)-Regulated Genes in Salmonella enterica Serovar Typhimurium. *J. Bacteriol.* 182: 5749-5756 [Abstract] [HTML]
- Bull, A. T., Ward, A. C., Goodfellow, M. (2000). Search and Discovery Strategies for Biotechnology: the Paradigm Shift. *Microbiol. Mol. Biol. Rev.* 64: 573-606 [Abstract] [HTML]
- McClelland, M., Florea, L., Sanderson, K., Clifton, S. W., Parkhill, J., Churcher, C., Dougan, G., Wilson, R. K., Miller, W. (2000). Comparison of the Escherichia coli K-12 genome with sampled genomes of a Klebsiella pneumoniae and three Salmonella enterica serovars, Typhimurium, Typhi and Paratyphi. *Nucleic Acids Res.* 28: 4974-4986 [Abstract] [HTML]